"Reversed" Faktor oder nicht? Ergebnisse einer subjektorientierten Reliabilitätsanalyse

Lettau, Frank

'Reversed' Faktor oder nicht?
Ergebnisse einer subjektorientierten Reliabilitätsanalyse

von Frank Lettau[1]

Abstract

In diesem Beitrag wird der Frage nachgegangen, inwieweit sich der bei der Verwendung von 'gedrehten' - sonst aber eindimensionalen - Items auftretende 'Negativ-Faktor' als eine Auswirkung mangelnder Subjekt-Reliabilität erklären läßt. Mangelnde Subjekt-Reliabilität wird dabei als Abweichung vom meßtheoretischen Idealzustand eindimensional-kumulativer Messung verstanden, wie er z.B. durch die Guttmannsche 'scalability' impliziert ist, und wie er explizit als Modellannahme in 'Strong true score'- Modelle, wie die Rasch-Responsemodelle eingebunden ist. Analysiert werden die Responses auf eine eindimensionale 10-Item-Skala, bei der drei der zehn Items 'gedreht' sind. Eine unter Anwendung eines generalisierten Rasch-Modells durchgeführte Reliabilitätsanalyse der Antwortmuster führt zu dem Ergebnis, daß ein Großteil dessen, was als Evidenz für eine zweidimensionale latente Struktur bzw. einen 'gedrehten' Faktor aufscheint, durch einige wenige unplausible Antwortmuster verursacht wird. Ganz allgemein bestätigt sich, daß den subjektbedingten Responsestörungen gerade bei der Messung latenter Variablen ein nicht zu unterschätzendes Gewicht innerhalb des Bereichs der 'Non-Sampling' -Surveyfehler zukommt.

The ' reversed factor' is a common phenomenon in social research: Responses to almost any set of items measuring a subjective variable tend to generale a ' reversed' factor or a bivariate latent structure, if some of the indicators are worded 'reversely'. Analyzing data consisting of responses to a 10-item-instrument measuring 'self-competenee' - out of which three items are 'reversed' - this paper investigates whether the apparent bivariate latent structure can be linked to a lack of person reliability which in turn may have been caused by non-attitudes, unwillingness to respond, carryover effects, misinterpretation or simply by missing the item reversals due to lacking attention etc. Person reliability is measured here as deviation from a generalized Rasch model, which - as a 'strong true score'-model - is a formulation of idealized measurement conditions with unidimensional cumulative item sets, additive conjoint measurement of the latent variable in question etc. Results show that the ' reversed factor' can infact be explained to a considerable extent by a lack of person reliability: Removal of only afew and evidently very unplausible responses leads to a much better fit for single-factor model. It can be concluded that much of the evidence for the presence of a 'reversed factor' may infact be caused by afew crude and unplausible response patterns.

---

1 Dr. Frank *Lettau* ist Wissenschaftlicher Assistent im Arbeitsbereich 'Methoden' des Instituts für Soziologie der TU Berlin, Dovestraße 1, W-1000 Berlin 10.

Negativ- oder 'reversed' Faktoren sind ein in der sozialwissenschaftlich-empirischen Forschung bekanntes Phänomen: Fast immer, wenn eine latente Dimension über mehrere Indikatoren gemessen wird und einige dieser Indikatoren dadurch gedreht ('reversed') sind, daß sie eine im Vergleich zum Rest der Indikatoren gegensätzliche Frageausrichtung aufweisen, tritt auch ein deutlich erkennbarer zweiter Faktor bzw. eine zweite Hauptkomponente auf, die überwiegend in Zusammenhang mit den 'reversed' Items steht.

Die Frage ist, wie ein solcher Sachverhalt zu interpretieren ist. Die wohl am häufigsten vorzufindende Sichtweise ist die, den 'reversed' Faktor als eine Art methodisches Artefakt zu betrachten, da sämtliche Items - auch die 'reversed' Items - letztlich als auf *eine* latente theoretische Dimension bezogen intendiert sind. Der 'reversed' Faktor wird demzufolge als das Resultat einer systematischen Antwortverzerrung - eines 'Response-Set' - angesehen, das durch die 'gedrehte' Formulierung der Itemtexte hervorgerufen wird. Die alternative Sichtweise wäre, positiv und negativ formulierte Items (a priori) als Indikatoren auch theoretisch eigenständiger latenter Dimensionen zu interpretieren [2].

Carmines & *Zeller* (1979:63f) gehen auf die Probleme im Zusammenhang mit der Verwendung von 'reversed' Items ein. Sie kommen durch Faktorenanalysen der zehn Items umfassenden 'Self-Esteem'-Skala von Morris *Rosenberg* (1965), die je fünf positiv formulierte und fünf negativ formulierte Items enthält, zu keinem eindeutigen Ergebnis (op.cit.: 67): '... since the bifactorial structure can be a function of a single theoretical dimension which is contaminated by a method artifact as well as being indicative of two separate, substantive dimensions, the factor analysis leaves the theoretical structure of self-esteem indeterminate.' Beim Umgang mit derartigen Skalen und der Erklärung der sich ergebenden Korrelationsstrukturen befindet sich der Forscher also in dem Dilemma, sich entweder auf das Vorhandensein eines nicht immer eindeutig begründbaren Response-Sets zu berufen, oder aber die Existenz einer weiteren - meist nicht intendierten - latenten Dimension zu akzeptieren [3].

---

[2] In einem solchen Fall sollten, strenggenommen, die einzelnen Subskalen auch als eigenständige Instrumente eingesetzt werden. Wie *Waltz* ((1987); vgl. hierzu auch *Pfaff* (1989)) bemerkt, ist insbesondere bei Affektskalen wie z.B der 'Affect Balance Scale' von *Bradburn* (1969), häufig von einer bidimensionalen latenten Struktur auszugehen.

[3] Im konkreten Anwendungsfall wird man natürlich versuchen, weitere Evidenz für die eine oder die andere Interpretationsmöglichkeit - insbesondere durch externe Validierung - beizubringen.

Vorgehensweise

In diesem Beitrag soll das Problem der 'reversed' Items anhand einer exemplarisch durchgeführten Datenanalyse erneut angegangen werden, diesmal jedoch aus dem Blickwinkel der *Reliabilität des individuellen Antwortverhaltens.* Als Arbeitshypothese wird der erstgenannte Standpunkt eingenommen, daß es für die Erfassung einer latenten theoretischen Dimension prinzipiell unerheblich sein sollte, ob die Items eines Instruments eine gemeinsame Frageausrichtung aufweisen oder nicht. Entgegenstehende Evidenz - insbesondere eine auf mehrere latente Dimensionen hindeutende Korrelations- und Faktorenstruktur - wird dabei als ein Artefakt betrachtet, das auf mangelnde Reliabilität des Antwortverhaltens bzw. mangelnde Personenreliabilität zurückzuführen ist. Diese äußert sich ganz allgemein in plausiblen Antwortmustern, die nicht kohärent mit den bei der Skalenkonstruktion zugrundegelegten Annahmen hinsichtlich Dimensionalität und Funktionsweise der Indikatoren sind. Response-Sets lassen sich in diesem Zusammenhang als eine der möglichen Manifestationen mangelnder Reliabilität des Antwortverhaltens unterordnen [4].

Es soll folgendermaßen vorgegangen werden: Als erstes werden das Erhebungsinstrument, die Erhebungsumstände und die Struktur der erhobenen Daten erläutert; im nächsten Schritt wird kurz auf das statistische Modell eingegangen, aufgrund dessen man zu einem Maßstab für die Reliabilitätsbewertung von Antwortmustern gelangt; im letzten Schritt wird eine Personenselektion vorgenommen, und zwar aufgrund der zuvor für jede Untersuchungsperson erfolgten Reliabilitätsbewertung des Antwortverhaltens. Vorab sei nur angemerkt, daß es sich bei der hier angewendeten Konzeption von Personenreliabilität um einen der Guttman'schen Einteilung in 'scalables' und 'unscalables' nicht unähnlichen Bewertungsmaßstab handelt. Ganz generell entspricht die hier durchgeführte subjektorientierte Reliabilitätsanalyse einer *zeilenweisen* (unit-centered) Analyse der Datenmatrix; sie ist damit quasi komplementär zur spaltenweisen (variable-centered) Item-Reliabilitätsanalyse: Entsprechend einer Eliminierung nicht zur Skala passender Items erfolgt hier eine Eliminierung unplausibler Antwortmuster und damit eine *Partitionierung der Personenstichprobe in reliable und nicht reliable Respondenten.* Es werden sodann die Auswirkungen dieser Selektionsmethode auf die Korrelationsstruktur und die unterliegende latente Struktur des Datenmaterials im Hinblick auf die eingangs erläuterten Dimensionalitätshypothesen analysiert.

---

4 Ein für Skalen mit 'reversed' Items typischer Response-Set dürfte sich dadurch konstituieren, daß ein Teil der Untersuchungspersonen die 'reversals' in den Schwerpunkten der Frageformulierungen übersieht.

Beschreibung des Datenmaterials

Das zu analysierende Datenmaterial umfaßt Responses auf zehn bewährte Indikatoren für die Messung der latenten Dimension 'Self-Competence' (vgl. ***Martimer & Lorence*** (1979); ***Quinn*** & ***Staines*** (1979)) [5]. Die Stichprobe besteht aus 1004 Schülern im Alter von 11-17 Jahren [6]. Diese Jugendlichen waren u.a. gefordert, sich im Hinblick auf ihre schulischen Leistungen und ihre schulische Betätigung insgesamt einzustufen, wozu ihnen das folgende Polaritätenprofil vorgelegt wurde:

Überlege einmal, wie Du Dich im Hinblick auf die Schule beurteilst. Denke dabei nicht nur an den Unterricht, sondern auch an Deine Klassenarbeiten, Deine Hausarbeiten, Deine Arbeitsgemeinschaften und an alles andere, was mit der Schule zu tun hat.

... Instruktion zum Bearbeiten eines Polaritätenprofils ...

Alles in allem beurteile ich mich folgendermaßen:
(Bitte in jeder Zeile ein Kreuz machen)

		1 2 3 4 5	
SC01	sehr erfolgreich	o----o----o----o----o	nicht allzu erfolgreich
SC02-R	schwach	o----o----o----o----o	stark
SC03-R	nicht allzu fähig	o----o----o----o----o	sehr fähig
SC04	zuversichtlich	o----o----o----o----o	besorgt
SC05	sehr aktiv	o----o----o----o----o	eher passiv
SC06	sehr tüchtig	o----o----o----o----o	nicht allzu tüchtig
SC07-R	nicht allzu sicher	o----o----o----o----o	sehr sicher
SC08	sehr zielstrebig	o----o----o----o----o	nicht allzu zielstrebig
SC09	kenne mich sehr gut aus	o----o----o----o----o	kenne mich nicht allzu gut aus
SC10	traue mir sehr viel zu	o----o----o----o----o	traue mir nicht allzu viel zu

Abb.l: Instrument zur Messung der Dimension 'Self-Competence'.

[5] Diese Items wurden von *Benninghaus* (vgl. 1980; 1987) in deutscher Übersetzung erfolgreich eingesetzt. Erfahrungsgemäß liefern diese Items recht schiefe Verteilungen, weshalb die Polaritäten hier mit Zusätzen wie *'sehr...'* an einem Ende und *'nicht allzu...'* oder *'eher...'* am anderen Ende des Konlinuums verschen wurden.

[6] Die Befragung erfolgte 1986 und ist Teil einer großen Längsschnittuntersuchung zum Bereich Jugendentwicklung (DFG Si 296/1-6; Projektleiter: R.K. *Silbereisen* und K. *Eyferth;* des weiteren IFP 2/11 der TU Berlin). Stichprobe: Geschichtete Zufallsauswahl.

Es sind zwei Item-Subgruppen zu erkennen: Eine aus sieben Items mit gemeinsamer Frageausrichtung bestehende Gruppe (nachfolgend 'G.7' genannt) und eine weitere Gruppe bestehend aus den drei 'R'-Items, die eine entgegengesetzte Frageausrichtung aufweisen ('G.3'). Die vorliegenden Daten sind so recodiert und aufbereitet, daß sich eine additive 10-Item-Skala ('SELFC') ergibt, bei der ein hoher Summenwert auch einer hohen Einschätzung des Selbstkompetenzgefühls entspricht. Als eine Möglichkeit zur externen Validierung dieser Skala wird ein Index für die tatsächliche schulische Leistung ('LEIST') mitberücksichtigt [7]. Um einen ersten Eindruck von der Datenstruktur zu gewinnen, seien zunächst die Korrelationen der Items untereinander sowie mit dem Leistungsindex präsentiert:

ITEM	01	04	05	06	08	09	10	02-R	03-R	07-R	SELFC
01	***										.489
04	.352	***									.411
05	.264	.335	***								.469
06	.270	.182	.330	***							.368
08	.312	.248	.369	.361	***						.528
09	.290	.248	.269	.242	.415	***					.497
10	.369	.308	.317	.266	.422	.509	***				.568
02-R	.272	.128	.176	.086	.184	.216	.190	***			.375
03-R	.201	.176	.194	.070	.217	.248	.294	.429	***		.422
07-R	.268	.238	.228	.166	.247	.202	.305	.393	.471	***	.465
LEIST	.315	.146	.160	.193	.175	.100	.153	.130	.181	.164	.291

Alpha-Reliabilität:.789; durchschnittliche Inter-Item-Korrelation:.273

Abb 2: Korrelationen, Reliabilitäts- und Validierungskennwerte für die Skala 'Self-Competence' [8].

Diese Kennwerte lassen das hier für die Messung der Dimension 'Self-Competence' eingesetzte Instrument unter Reliabilitäts- und Validitätsaspekten als insgesamt recht brauchbar erscheinen: Die Skala weist zufriedenstellende Grade an interner Konsistenz und externer Validität auf. Auf eine Besonderheit soll hier nur hingewiesen werden, nämlich den relativ starken Zusammenhang des ersten Items **SC01** des Polaritätenprofils mit dem

---

[7] Der Index ist als Summe der Noten für die Schulfächer Deutsch und Mathematik konstruiert und so recodiert, daß hohe schulische Leistung einem hohen Indexwert entspricht

[8] Die Zahlenwerte unter 'SELFC repräsentieren die 'corrected item-total correlations' der einzelnen Indikatoren mit dem Summenindex SELFC.

Validierungskriterium **LEIST**[9]. Was den möglichen Zusammenhang von latenter Dimensionalität und unterschiedlicher externer Validierung anbetrifft, so ist erkennbar, daß die Indikatoren beider Itemgruppen ähnlich stark mit **LEIST** korrelieren: Die durchschnittliche Korrelation beträgt für G.7 0.177 bzw. 0.155 bei Nichtberücksichtigung des sehr starken 'Eingangsitems' **SC01;** für **G.3** beträgt diese 0.158. Diese dicht beieinanderliegenden Werte legen die Annahme einer bidimensionalen latenten Struktur nicht zwingend nahe. Die entsprechend den Zugehörigkeiten der Items zu G.7 und G.3 umgeordnete Korrelationsmatrix der Indikatoren läßt bei genauerem Hinsehen allerdings eine Struktur erkennen, in der sich beide Itemgruppen deutlich gegeneinander differenzieren. Diese Struktur tritt klarer zutage, wenn man die durchschnittlichen Korrelationen *innerhalb* der Itemgruppen G.7 und G.3 sowie *zwischen* diesen vergleicht:

	G.7	G.3
G.7	0.318	
G.3	0.205	0.431

ausgedrückt in Verhältniszahlen zum Durchschnitt der Korrelationen:

	G.7	G.3
G.7	1.093	
G.3	.751	1.579

Diese für Skalen mit gedrehten Items nicht untypische Korrelationsstruktur mit abgeschwächten Korrelationen zwischen den Itemgruppen läßt eher die Annahme der Existenz zweier separater latenter Dimensionen gerechtfertigt erscheinen.

An diesem Punkt angelangt, könnte man sagen, daß sich Evidenz sowohl für die latent-eindimensionale wie auch die latent-zweidimensionale Interpretation der Daten ergibt. Um zu einer vergleichenden Bewertung beider Dimensionalitätshypothesen zu gelangen, werden beide als Modelle der konfirmatorischen Faktorenanalyse formuliert und überprüft [10]:

---

9 Die hier aufzuwerfende Frage wäre, ob es sich um einen Plazierungseffekt handelt (SC01 als 'Eingangsitem' des Polarilätenprofils, bei dem die Instruktionen zum Ankreuzen noch präsent sind), oder ob eine inhaltliche ('item content') Begründung zutreffender ist Zur Überprüfung dieser Frage (und generell zur Ausschaltung systematischer Fehlerquellen) wäre eine Randomisierung der Itemsequenzen pro Instrument sinnvoll.

10 ML-Schätzwerte, die mit Hilfe des LISREL-Programms (vgl. *Jöreskog & Sörbom,* 1981) in Anpassung an die Korrelationsmatrix ermittelt wurden. Eine Maximierung in Anpassung an die Kovarianzmatrix, wie sie von den Schätzgrundlagen her (Maximierung des Likelihood unter Annahme der Wishart-Verteilung für Kovarianzmatrizen) angemessener wäre, führt zu exakt übereinstimmenden Lösungen sowohl hinsichtlich Modell-Fit, wie auch - nach entsprechender Umrechnung der Faktorenladungen in Kommunalitäten - hinsichtlich der Schätzwerte.

Modell (M-1D) der Eindimensionalitätshypothese

$$\begin{bmatrix} \hat{G.7} \\ \hline \hat{G.3} \end{bmatrix} = \begin{bmatrix} \lambda \\ \hline \lambda \end{bmatrix} * \begin{bmatrix} D \end{bmatrix}$$

Modell (M-2D) der Zweidimensionalitätshypothese

$$\begin{bmatrix} \hat{G.7} \\ \hline \hat{G.3} \end{bmatrix} = \begin{bmatrix} \lambda & - \\ \hline - & \lambda \end{bmatrix} * \begin{bmatrix} D1 \\ \hline D2 \end{bmatrix}$$ D1, D2 korreliert

Faktorenladungen (λ-Matrix)

(M-1D)	(M-2D)		
.546	.542	-	
.469	.473	-	
.520	.530	-	
.428	.453		
.611	.634	-	D1
.597	.618	-	
.673	.690	-	0.561
.409	-	.591	
.463	-	.693	
.500	-	.686	D2

Modell-Fit

(M-1D)	(M-2D)
Chi^2	
(df=35)	(df=34)
459.90	186.33
Chi^2/df	
13.14	5.48

Abb. 3: Konfirmatorische Faktorenanalysen: Schematische Darstellung der Modelle sowie der entsprechenden LISREL-Schätzergebnisse.

Bei einer Betrachtung der Faktorenladungen zeigt sich das von Skalen mit 'reversed' Items bekannte Bild einer zweidimensionalen latenten Struktur, welche sich bei Inspektion der Korrelationsmatrix ja bereits angedeutet hatte. Der Eindruck, daß das Postulat einer zweidimensionalen latenten Struktur mit den vorliegenden Daten wesentlich besser verträglich ist, bestätigt sich durch die weit auseinanderliegenden Chi^2-Kennwerte mit einer Chi^2-Differenz von 273.57 [11]. Wie an der recht hohen geschätzten Korrelation beider latenten Dimensionen ersichtlich ist, bedeutet eine bessere Anpassung des latent-zweidimensionalen Modells jedoch nicht, daß beide latenten Dimensionen inhaltlich nicht dicht beieinander liegen könnten.

11 Dieser Chi^2 kann als Teststatistik mit (df=l) interpretiert werden, da es sich um 'nested models' handelt: M-1D läßt sich auch als zweifaktorielles Modell deuten, deren latente Dimensionen Dl und D2 mit 1.0 korreliert sind. M-1D entspricht insofern einer mit einer zusätzlichen Restriktion versehenen Fassung von M-2D.

Reliabilitätsbetrachtung der Antwortmuster

Die Frage, die sich jetzt stellt, ist, ob die das bifaktorielle Erscheinungsbild hervorrufende Differenzierung des Antwortverhaltens nur ein Artefakt ist, welches sich auf personenbezogene Störungen des Antwortverhaltens gründet. Wenn dies der Fall ist, dann müßte sich die faktorielle Differenzierung durch eine personenorientierte Reliabilitätsanalyse und die damit einhergehende Eliminierung der unplausibelsten Antwortmuster zum Verschwinden bringen lassen. Zur Illustration seien hier einige derjenigen Antwortmuster vorgeführt, die man als unplausibel bzw. indikativ für mangelnde Personenreliabilität bezeichnen könnte [12]:

	Untersuchungsperson							
Itemtext	017	041	161	336	364	446	564	677
...erfolgreich.....	3	5	5	5	4	4	1	1
...stark(r).....	2	5	1	1	3	5	1	5
...fähig(r).....	5	5	1	1	2	5	5	5
...zuversichtlich...	3	1	5	5	2	3	1	1
...aktiv.....	5	5	5	5	1	4	5	5
...tüchtig.....	1	5	5	5	5	5	5	5
...sicher(r).....	5	5	1	1	1	5	1	1
...zielstrebig.....	5	5	5	4	1	1	5	5
...gut auskennen....	5	5	5	3	5	1	5	5
...viel zutrauen....	5	5	5	2	1	1	1	5

Abb. 4: Einige Beispiele für unplausible Antwortmuster bei der Skala 'SELFC

Die mangelnde Plausibilität dieser Antwortmuster ist relativ leicht zu erkennen; ohne hier auf einzelne Antwortmuster einzugehen, lassen sich überschlägig die folgenden Responsefehler feststellen: a) Übersehen der 'item reversals' (#161, #336, #677); b) einzelne im Vergleich zum sonstigen Antwortverhalten unerwartete Subjekt-Item-Interaktionen (#041); c) uniformes Antwortverhalten entlang einer Seite bzw. einer Spalte des Polaritätenprofils (#161); d) absolut erratische Antwortmuster, die nur durch völliges Mißverstehen, Konzentrationsmängel oder einfach fehlende Motivation zu erklären sind (#564) und - als

---

[12] Die Kodierung der Responses entspricht nicht mehr der in Abb.l; vielmehr sind die Responses jetzt zwecks besserer Lesbarkeit einheitlich so recodiert, daß ein hoher Itemscore einer hohen 'Self-Competence'-Einschätzung entspricht.

letztes - Kombinationen dieser Möglichkeiten (#017, #336, #364, #446, #677) [13]. Eine solche ins Detail gehende Inspektion einzelner Antwortmuster ('data editing') kann wichtige Hinweise auf mögliche Fehlerquellen bei der Datenerhebung liefern [14]. Es liegt auf der Hand, daß wenn man grobe und offensichtliche Fehlbeantwortungen in die Datenanalyse einbezieht, diese dort einen mehr oder weniger starken 'noise' verursachen, der das Erkennen von Daten strukturen potentiell erschwert.

Es ist nun in einem kurzen Exkurs darauf einzugehen, nach welchem Prinzip bzw. mit Hilfe welcher Methodik hier eine Reliabilitätsbewertung der einzelnen Antwortmuster erfolgt. Theoretisch wäre der Versuch denkbar, unreliable Respondenten durch eine genaue Inspektion der Rohdatenmatrix zu entdecken - es sollten sich auf diese Weise zumindest einige der grob unplausiblen und besonders auffälligen Antwortmuster aufspüren lassen. Diese 'Eye-Ball'-Methode hätte zwei Nachteile: Zum einen stieße sie bei einigen hundert oder gar tausend Fällen an eine Grenze des technisch Machbaren, zum anderen wäre eine solche Selektion auch deswegen 'unsauber' bzw. theoretisch unbefriedigend, weil sie nach Ad hoc-Kriterien erfolgte und nicht auf einem expliziten und intersubjektiv gültigen Standard beruhte. Deswegen wird hier ein anderer Weg eingeschlagen - und zwar wird die Plausibilität des Antwortverhaltens einer jeden Person anhand eines statistischen Modells bewertet, welches den psychologischen Prozeß der Beantwortung eines Rating-Items nachbildet und daraus dann konkrete Antwortwahrscheinlichkeiten ableitet. Dieses Modell ist das sog. '*Rating-Response-Modell*` (nachfolgend 'RRM'); es ist andernorts ausführlich dokumentiert (vgl. dazu z.B. *Andrich* (1978a, 1978b, 1978c, 1979), *Wright & Masters (1982), Lettau* (1987), *Rost* et al. (1990)) und wird deshalb an dieser Stelle nur so weit

---

13 Bei genauerer Inspektion der Responsemuster lassen sich evtl. sogar Tendenzen einer Zu- oder Abnahme der Personenreliabilität erkennen: Während Person #336 am Ende der Itemsequenz ein besseres (differenziertes) Response-Verhalten als zu Beginn zeigt, offenbaren die Personen #017, #364 und #446 ein gegen Ende zunehmend erratischeres Responseverhalten.

14 Eine vollständige Aufarbeitung möglicher Ursachen von subjektbedingten Responsefehlern würde den Rahmen dieses Beitrages sprengen. Neben dem schon genannten und nur in diesem speziellen Zusammenhang wichtigen Übersehen der 'item reversals' dürften die meisten Response-Störungen eher unsystematischer und unregelmäßiger Natur sein, wie etwa Konzentrationsfehler, spezifische Subjekt/-Item-Interaktionen (besondere Bedeutungsbeilegungen oder Mißverstehen bestimmter Items) und kurzfristige Fluktuationen der Einstellungsvariablen während der Messung (sog. 'attitude tremors'; vgl. z.B. *Lumsden* (1977, 1978)). Von besonderer Tragweite für eine Diskussion möglicher Fehlerquellen bei der Einstellungsmessung ist die - hauptsächlich durch Philip *Converse* (vgl. 1964, 1970) ins allgemeine Bewußtsein gehobene - Problematik der sog. 'Non-Attitudes' bzw. spezieller: der 'hidden Non-Attitudes' (vgl. z.B. die Zusammenfassung von *Smith* (1984, speziell ab S. 230), oder auch *Reuband* (1990)). Es läßt sich vermuten, daß die vorliegenden Daten ein erhebliches Maß aller denkbaren Arten von 'noise' enthalten, da a) dieses Instrument erst gegen Ende der insgesamt etwa zwei bis drei Stunden (!) dauernden Befragung durchlaufen wurde; b) die Stichprobe aus Kindern und Jugendlichen besteht und c) ein Polaritätenprofil hier zum ersten Mal im Laufe der Erhebung eingesetzt wird.

erläutert, daß zum einen der zugrundegelegte Rating-Mechanismus veranschaulicht wird, und zum anderen sich der Zusammenhang mit der hier verfolgten Reliabilitäts-Fragestellung herstellen läßt.

Das **RRM** gehört zur Gruppe der Rasch-Modelle bzw. Logistischen Latent-Trait-Modelle (LLTM), deren Kern das sog. Einfache Logistische Modell (SLM) ist (vgl. ***Rasch*** (1960, 1966); ***Wright & Stone*** (1979)) [15]. Ganz allgemein werden die Rasch-Modelle zur sog. *'Strong true score'-Theorie* gerechnet, weil sie den bei Zusammentreffen eines Items und einer Untersuchungsperson ablaufenden Responseprozeß explizit ausformulieren. Hierzu wiederum bedarf es einiger vereinfachender Annahmen, die einer Umschreibung meßtheoretisch idealer Bedingungen nahekommen [16]. Dies betrifft zunächst die Annahme der *Eindimensionalität:* Sowohl für die Items wie auch für die Untersuchungspersonen gilt, daß die einzig relevante Unterscheidung zwischen ihnen in ihrer Plazierung auf dem Kontinuum der zu messendenden latenten Dimension besteht - etwa in dem Sinn, daß eine Zustimmung zu einem Item 'A' höhere 'self-competence' erfordert als die Zustimmung zu einem Item 'B'. Item 'A' ist damit höher auf dem latenten Kontinuum zu verorten als Item 'B'. Entsprechendes gilt für die Untersuchungspersonen. Diese Lozierungsparameter für Items ('item-difficulty' oder (δ)-Parameter) und Personen ('person ability' oder β-Parameter) sind die zentralen Parameter aller Rasch-Modelle [17].

Eine weitere wichtige Annahme der LLTM besteht darin, daß das Zusammentreffen eines Items **i** (mit einer Plazierung δ_i auf dem latenten Kontinuum) und einer Untersuchungsperson v (mit einer Plazierung β_v) als ein einfacher *additiver Prozeß* $(\beta_v-\delta_i)$ erfolgt, der bei allen Items, bei allen Untersuchungspersonen und in allen Bereichen des latenten Kontinuums auf gleiche Weise und ohne jegliche Interaktionseffekte wirksam wird, womit auch in diesem Punkt von einer idealen Meßsituation ('additive conjoint measurement';

---

15 Das SLM ist ein Responsemodell für dichotome Items, während die anderen LLTM wie das RRM Adaptionen des SLM für polytome Antwortformate darstellen. *Masters* & *Wright* (1984) geben eine Übersicht verschiedener Ableitungen des SLM (vgl. auch ***Andrich*** (1982); *Rost* et al. (1990)).

16 ***Duncan*** (1984a, 1984b) geht auf die den LLTM zugrundeliegenden 'Strong true score'-Annahmen ein und erläutert deren Relevanz für den Kontext der Umfrageforschung.

17 Die erschöpfenden Statistiken für den δ- und den β-Parameter sind die jeweiligen Summenwerte der Items und Untersuchungspersonen. Wie die Guttman- oder die Likert-Skalierung werten damit auch Rasch-Modelle nicht das 'Innere' einer Datenmatrix aus, sondern nur die Spaltensummen **(δ-Parameter)** und die Zeilensummen **(β-Parameter).** Diesen Skalierungsmethoden liegt die gemeinsame Annahme zugrunde, daß eine monotone Beziehung zwischen dem Wert des Summenindex und der (latenten) Item- oder Personeneinstufung besteht. Die Rasch-Modelle unterscheiden sich von den vorgenannten Verfahren dadurch, daß sie die einfachen Summenwerte in 'bessere' Schätzwerte (stichprobenunabhängige und ML-optimierte Logits) für die Item- und Personeneinstufungen liefern (vgl. für Einzelheiten der (UCON-) Schätzung z.B. ***Wright*** & *Panchapakesan* (1969); *Wright* & *Douglas* (1977); *Wright* & *Masters* (1982)).

vgl. ***Perline*** et al. (1979), ***Wright*** (1984)) ausgegangen wird. Betrachtet man als letztes die aus der Einbindung dieser Annahmen in eine logistische Funktion resultierenden Antwortwahrscheinlichkeiten (vgl. hierzu die Abb. im Anhang), so wird deutlich, daß Rasch-Modelle im großen und ganzen eine Umsetzung dessen sind, was üblicherweise bei der Konstruktion und Anwendung eindimensionaler Skalen als grundlegende Monotonieeigenschaft vorausgesetzt wird: Je höher eine Person die eigene 'self-competence' einschätzt (vgl. Person v_2 mit der hohen Einstufung $ß_2$), desto größer ist auch die Wahrscheinlichkeit, dieses Item - je nach Antwortformat unterschiedlich - zustimmend zu beantworten und umgekehrt (vgl. Person v, mit der niedrigen Einstufung $ß_t$).

Rasch-Skalierungsmodelle sind im Hinblick auf die erwähnten Annahmen bezüglich Eindimensionalität und Monotonieeigenschaft, über die sich die Verwendung von Zeilen- oder Spaltensummen als Repräsentationen eines latenten Kontinuums rechtfertigen läßt, mit der Guttman- oder der Likert-Skalierung vergleichbar (s. Fußnote 17). Was die Rasch-Modelle aber von diesen unterscheidet und im Zusammenhang mit der Frage der Reliabilitätsbewertung besonders brauchbar macht, ist, daß sie aufgrund ihrer probabilistischen Konzeption keine starre Einteilung der Personenstichprobe in 'scalables' oder 'unscalables' kennen. Stattdessen liefern diese Modelle *aus den Antwortwahrscheinlichkeiten* $p(x_{vj}=0,l,2,3,...)$ *abgeleitete Erwartungswerte* $E(x_{vi})$ für jede einzelne Response x_{vi}, wodurch sich ein Vergleich von tatsächlichem und erwartetem Antwortverhalten anstellen läßt. Aus diesen Einzel-Residuen $(x_{vi}-E(x_{vi}))=Res(x_{vi})$ wiederum läßt sich ein summarischer Index für die Beurteilung der Personenreliabilität dadurch gewinnen, daß man diese über sämtliche Items der Skala hinweg zeilenweise zusammenfaßt - also die $Res(x_{v+})$ bildet [18]. Es ergibt sich so eine Beurteilungsmöglichkeit für den Grad der 'scalability' des *gesamten Antwortmusters* eines Respondenten. Die Anpassung der Antwortvektoren der in Abb. 4 aufgeführten Untersuchungspersonen #017, #041 und #161 an das RRM stellt sich bspw. folgendermaßen dar:

	Abweichungen $Res(x_{vi}) \geq 1.0$ für die Items SC01-SC10										Reliabilitätsindex $Res(x_{v.})$
v= i=	01	02r	03r	04	05	06	07r	08	09	10	
#017:	-1.1	-2.6	1.4	-1.3	1.3	-3.6	1.6	1.5	1.5	1.4	3.5
#041:	*	*	*	-6.7	*	*	*	*	*	*	5.0
#161	1.7	-3.8	-4.0	1.5	1.4	1.9	-3.5	1.7	1.6	1.5	6.1

[18] Die $Res(x_{v.})$ sind hier einer Slandardisierungtransformation unterworfen, die den Effekt hat, die Abweichungen $Res(x_{vi})$ bei großem Abstand $(\beta_v-\delta_i)$ stärker zu gewichten. Die summarische Anpassungsstatistik $Res(x_{v.})$ ist ganz einfach das mittlere quadratische Residuum; dessen Erwartungswert liegt bei 1.0 (vgl. z.B. *Wright* & *Masters* (1982:94f.) für eine ausführliche Darstellung weiterer Details über Fit-Indices für LLTM).

Erkennbar wird hierbei, wie individuell verschieden die Abweichungen vom Standard dieses eindimensionalen 'strong true score'-Modells ausfallen: Person #017 legt ein über alle Items verteiltes, etwas diffuses Antwortverhalten an den Tag, Person #041 'leistet' sich eine drastische Fehlbeantwortung, und die Anpassung von Person #161 leidet - wie hier offensichtlich wird - unter dem Übersehen der Itemdrehungen. Unabhängig von der Art der Fehlbeantwortungen ergeben sich in jedem Fall schlechte Werte für die Gesamtanpassung. Das RRM signalisiert auf diese Weise deutlich die mangelnde Reliabilität bzw. die grobe Unverträglichkeit der vorliegenden Antwortmuster mit dem Modell einer eindimensional-kumulativen Skala.

Ergebnisse

Es sollen die Auswirkungen auf die Datenstruktur beschrieben werden, die sich bei einer Bereinigung der Daten um die am wenigsten reliablen Antwortmuster ergeben. Es erscheint sinnvoll, zunächst die Anpassung aller N=1004 Respondenten bzw. ihrer Antwortmuster an das RRM zu betrachten. Es ergeben sich hieraus Anhaltspunkte für die vorzunehmende Partitionierung der Stichprobe in reliable und unreliable Respondenten:

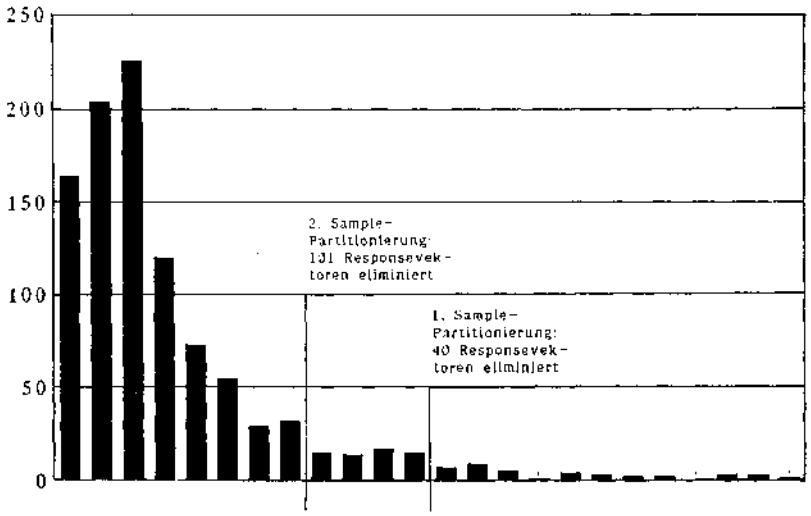


Abb. 5: Verteilung der Personenreliabilität gemessen durch die Anpassung $\mathbf{Res}(\mathbf{x}_{v+})$ an das Rating-Response-Modell ($\bar{x}$= 1.000; s=0.984)

Wie sich an der Schiefe dieser Verteilung zeigt, sind es nur relativ wenige Personen, die die Items auf eine Weise bearbeiten, die der mit Hilfe des RRM operationalisierten Erwartung hinsichtlich eines kohärenten Antwortverhaltens grob zuwiderläuft. Eingangs war bereits erwähnt, daß hier keine anderen als nur diese sehr unplausiblen Antwortmuster eliminiert werden sollen; das RRM dient damit nur als Hilfsmittel zur Indizierung wirklich krasser Fehlbeantwortungen und wird sehr behutsam als ein Standard für 'richtiges' Antwortverhalten auf die Daten angewendet. Wie in Abb. 5 angedeutet, werden hier zwei Partitionierungen der Stichprobe vorgenommen: Die erste stellt nur einen sehr leichten Eingriff in die Stichprobe dar - durch sie werden diejenigen Respondenten aus dem Sample entfernt, deren $\mathbf{Res}(x_{v*})$ mehr als etwa 2.25 Standardabweichungen vom Mittelwert 1.0 entfernt ist; hierbei handelt es sich um 40 Fälle, zu denen auch die in Abb. 4 dargestellten Antwortvektoren gehören. Durch die zweite Partitionierung werden weitere 61 wenig reliable Respondenten aus der Stichprobe entfernt, wodurch deren Umfang von N=1004 auf N=903 sinkt. Diese zweite Partitionierung entspricht damit bereits einer etwas nachdrücklicheren Durchsetzung des RRM als Standard für skalenadäquates Antwortverhalten.

Als erstes seien jetzt die Auswirkungen auf die Korrelationsstruktur, die Reliabilitäts- und Validitätskennwerte und die faktorielle Struktur des Itemsatzes betrachtet, die sich nach der ersten Partitionierung durch Weglassung der 40 unplausibelsten Antwortmuster ergeben:

ITEM	01	04	05	06	08	09	10	02-R	03-R	07-R	SELFC
01	***										.515
04	.328	***									.431
05	.267	.354	***								.501
06	.305	.195	.321	***							.410
08	.350	.265	.374	.366	***						.563
09	.305	.276	.310	.283	.447	***					.533
10	.375	.317	.358	.307	.443	.500	***				.582
02-R	.322	.167	.236	.164	.267	.268	.239	***			.433
03-R	.297	.253	.252	.133	.282	.292	.336	.412	***		.481
07-R	.320	.275	.301	.232	.311	.271	.328	.399	.463	***	.516
LEIST	.319	.141	.156	.218	.185	.124	.164	.143	.197	.166	.294

Alpha-Reliabilität:.815; durchschnittliche Inter-Item-Korrelation:.308

Abb 6: Korrelationen, Reliabilitäts- und Validierungskennwerte für die Skala 'Self-Competence' (1. Sample-Partitionierung - N=964)

Betrachtet man dazu noch die nach den beiden Itemsubgruppen G.7 und G.3 aggregierte Korrelationsmatrix,

	G.7	G.3
G.7	0.336	
G.3	0.264	0.425

ausgedrückt in Verhältniszahlen zum Durchschnitt der Korrelationen:

	G.7	G.3
G.7	1.091	
G.3	.857	1.380

so fällt die alles in allem höhere innere Konsistenz der 'Self-Competence'-Skala auf: a) die einzelnen Korrelationen sind im Durchschnitt recht deutlich erhöht; b) die an den Itemsubgruppen sichtbar gewordene und durch die Itemdrehungen verursachte Strukturiertheit der Korrelationsmatrix hat sich abgeschwächt, die Korrelationen sind wesentlich homogener geworden; c) die Item-Total-Korrelationen bzw. Item-Diskriminierungen sind höher als vorher und vor allem auch - bezogen auf die **G.7/G.3** Unterscheidung - homogener: Im Gesamtsample betrugen die durchschnittlichen Item-Total-Korrelationen 0.476 für G.7 und 0.421 für G.3; nach der ersten Eliminierung nicht reliabler Respondenten nähern sich diese Werte einander an - 0.505 für G.7 vs. 0.477 für G.3. Das Muster und die Stärke der Beziehungen zum externen Validierungskriterium **LEIST** werden durch die vorgenommene Partitionierung kaum beeinflußt: Das Eingangsitem **SC01** bleibt stark korreliert, während sich die einzelnen Korrelationen der Indikatoren mit **LEIST** im Durchschnitt ganz leicht erhöhen. Es entsteht insgesamt der Eindruck, daß sich die in den Ausgangsdaten gegebene Evidenz für eine bifaktorielle latente Struktur abgeschwächt hat. Dies wird durch die Ergebnisse der konfirmatorischen Faktorenanalyse unterstützt:

Faktorenladungen (λ-Matrix)

(M-1D)	(M-2D)		
.566	.560	-	D1
.480	.483	-	
.548	.556	-	
.463	.483		
.636	.657	-	
.612	.633	-	
.665	.686	-	
			0.697
.482	-	.592	
.536	-	.673	
.564	-	.691	D2

Modell-Fit

(M-1D)	(M-2D)
Chi^2	
(df=35)	(df=34)
285.65	129.57
Chi^2/df	
8.16	3.81

Abb. 7: Konfirmatorische Faktorenanalysen: Vergleichende Schätzung der Modelle M-1D und M-2D nach der 1. Sample-Partitionierung

Neben den im Sinne einer Annäherung an 'tau-equivalence' verbesserten Faktorenladungen des einfaktoriellen Modells erscheinen zwei Aspekte erwähnenswert: Dies ist zum einen die Korrelation der latenten Dimensionen mit einem geschätzten Wert von annähernd 0.70; und zum anderen ist dies die Fit-Verbesserung für die Eindimensionalitäts-Hypothese. Die Korrelationsmatrix ist zwar nach wie vor besser mit der Annahme einer bifaktoriellen latenten Struktur verträglich, jedoch hat sich die Verbesserung der Anpassung beim Übergang vom eindimensionalen zum zweidimensionalen Modell von ursprünglich Chi^2=273.57 auf jetzt Chi^2=156.08 verringert.

Die Auswirkungen der zweiten Stichprobenpartitionierung (Eliminierung von insgesamt 101 unreliablen Antwortvektoren) brauchen nicht im Detail kommentiert zu werden - vom Gesamteindruck verstärken sich die bisher schon gesehenen Veränderungstendenzen der Datenstruktur weiter:

ITEM	01	04	05	06	08	09	10	02-R	03-R	07-R	SELFC
01	***										.567
04	.358	***									.494
05	.332	.405	***								.547
06	.342	.217	.343	***							.427
08	.416	.337	.413	.387	***						.608
09	.349	.329	.353	.315	.481	***					.578
10	.395	.361	.393	.285	.480	.536	***				.604
02-R	.371	.218	.278	.221	.308	.296	.302	***			.477
03-R	.361	.325	.293	.144	.315	.323	.352	.422	***		.508
07-R	.385	.361	.368	.278	.378	.374	.398	.431	.480	***	.591
LEIST	.344	.173	.153	.231	.205	.140	.166	.144	.187	.190	.299

Alpha-Reliabilität:.843; durchschnittliche Inter-Item-Korrelation:.351

Abb 8: Korrelationen, Reliabilitäts- und Validierungskennwerte für die Skala 'Self-Competence' (2. Sample-Partitionierung - N=903)

Aggregiert nach den Itemgruppen G.7 und G.3 stellt sich die Korrelationsstruktur folgendermaßen dar:

G.7	0.373	
G.3	0.317	0.444
	G.7	G.3

ausgedrückt in Verhältniszahlen zum Durchschnitt der Korrelationen:

G.7	1.063	
G.3	.903	1.265
	G.7	G.3

Auch die faktorielle Struktur der zehn Indikatoren verschiebt sich weiter in Richtung auf eine verbesserte Anpassung des einfaktoriellen Modells und eine Erhöhung der geschätzten Korrelation der latenten Dimensionen im zweifaktoriellen Modell:

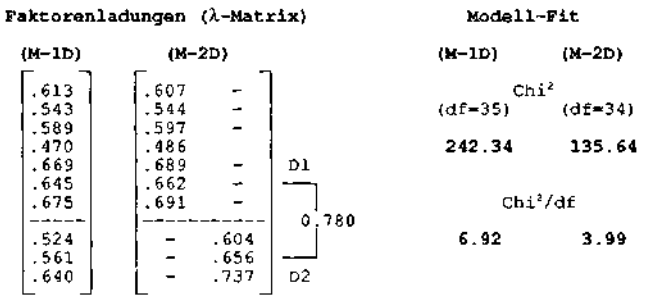


Abb. 9: Konfirmatorische Faktorenanalysen: Vergleichende Schätzung der Modelle M-1D und M-2D nach der 2. Sample-Partitionierung

Es fällt auf, daß sich - anders als bei der ersten Stichprobenpartitionierung - durch diese zweite und etwas 'schärfere' Partitionierung keine weitere Verbesserung der Anpassung der Zweidimensionalitätshypothese an die 'beobachteten' Korrelationen ergibt, während sich die Anpassung der Eindimensionalitätshypothese weiter verbessert, womit sich beide Modelle in Bezug auf ihre Datenanpassung weiter annähern (Fit-Differenz: Chi^2=106.70).

Zusammenfassung

Einmal abgesehen von einer im Einzelfall denkbaren theoretischen Rechtfertigung für die Annahme bifaktorieller latenter Strukturen zeigt sich hier empirisch, daß ein nicht unerheblicher Teil der für die Annahme einer bifaktoriellen latenten Struktur sprechenden Evidenz durch nur relativ wenige Untersuchungspersonen hervorgerufen wird. Betrachtet man die Ergebnisse insbesondere der sehr vorsichtigen ersten Stichprobenpartitionierung, so zeigt sich, daß sich neben einer allgemeinen Reduktion von 'noise' (erkennbar an insgesamt höheren Inter-Item-Korrelationen, besseren Anpassungswerten für beide Faktorenmodelle und geringfügig erhöhten Validierungskorrelationen) insbesondere eine 'noise'-Reduktion im Hinblick auf die Heterogenität der Itemgruppen G.7 und G.3 ergibt.

Es deutet sich damit an, daß ein erheblicher Teil dessen, was in empirischen Untersuchungen als eine stark kontrastierende bifaktorielle Struktur aufscheint, möglicherweise durch eine systematische Antwortverzerrung, nämlich ganz einfach das 'Übersehen' der Itemdre-

hungen, ausgelöst wird. Der Verdacht, daß es sich beim Phänomen des 'reversed' Faktor um ein Artefakt handelt, wäre somit - zumindest teilweise - gerechtfertigt. Dieser Eindruck verfestigt sich, wenn man zum Vergleich die Korrelationsstruktur der nach der ersten Partitionierung entstandenen Unterstichprobe von N=40 ausselektierten Antwortvektoren betrachtet:

ITEM	01	04	05	06	08	09	10	02-R	03-R	07-R	SELFC
01	***										.274
04	.522	***									.189
05	.229	.176	***								.100
06	.002	.065	.355	***							-.095
08	.028	.115	.303	.287	***						.146
09	.170	.025	-.093	-.137	.152	***					.123
10	.328	.244	.017	-.029	.278	.564	***				.502
02-R-	.011	-.149	-.233	-.412	-.357	-.102	-.101	***			-.128
03-R-	.332	-.345	-.159	-.267	-.161	.025	.074	.522	***		-.027
07-R-	.017	-.017	-.241	-.213	-.139	-.205	.183	.342	.505	***	.063
LEIST	.340	.212	.248	.005	.032	-.114	.077	-.002	.022	.150	.282

Alpha-Reliabilität:.301; **durchschnittliche Inter-Item-Korrelation**:.040

Abb 10: Korrelationswerte der SELFC-Indikatoren für die Teilstichprobe der im Zuge der 1. Partitionierung ausselektierten Respondenten (N=40)

bzw. aggregiert nach den Itemgruppen G.7 und G.3:

G.7	0.171	
G.3	-0.151	0.456
	G.7	G.3

Stellt man diese Ergebnisse in einen allgemeineren Zusammenhang, so ließe sich fragen, ob die Methodik der Verwendung von 'reversed' Items, auf die in den meisten Fällen ohne zwingende inhaltliche Notwendigkeit als eine Technik zur Vermeidung von Monotonie-Responsesets zurückgegriffen wird, nicht mehr Schaden als Nutzen stiftet. Immerhin erfordert jede Art von Wechsel im Responseformat eine über die Erfassung des jeweiligen 'item content' hinausgehende Konzentrationsleistung, die je nach Erhebungsumständen nicht immer zu erbringen sein dürfte.

Das für eine Verwendung von 'reversed' Items sprechende Gegenargument könnte lauten, daß die Miteinbeziehung solcher Indikatoren gerade wegen der höheren zu erbringenden Konzentrationsleistung sinnvoll ist. Dieses Argument würde auf die möglichen Ursachen

für das falsche Bearbeiten von Itembatterien rekurrieren: Wenn es der Fall ist, daß ein Teil der Responses eines Datensatzes auf Unaufmerksamkeit, Demotiviertheit, Unverständnis des Instruments oder versteckte Meinungslosigkeit [19] zurückzuführen ist, dann könnte es sinnvoll erscheinen, gedrehte Items als 'Stolpersteine' mit in ein Instrument einzubeziehen, da diese unintendierten Responsefaktoren bzw. Response-Sets dann eher manifest werden und sich die entsprechenden Antwortmuster deutlicher von den erwartungskonformen Responses abheben.

Literatur

Andrich, D. (1978a): Scaling attitude items constructed and scored in the Likert tradition, Educational and Psychological Measurement, 38, 665-680.

Andrich, D. (1978b): Application of a psychometric rating model to ordered categories which are scored with successive integcrs, Applied Psychological Measurement, 2, 581-594.

Andrich, D. (1978c): A rating formulation for ordered response categories, Psychometrika, 43, 561-573.

Andrich, D. (1979): A model for contingency tables having ordered classifications, Biometrics, 35, 403-415.

Andrich, D. (1982): An extension of the Rasch model for ratings providing both location and dispersion Parameters, Psychometrika, 47, 105-113.

Benninghaus, H. (1980): Fragebogen 'Merkmale und Auswirkungen beruflicher Tätigkeit 1980', Seminar für Soziologie der Universität zu Köln.

Benninghaus, H. (1987): Substantielle Komplexität der Arbeit als zentrale Dimension der Jobstruktur, Zeitschrift für Soziologie, 16, 334-352.

Bradburn, N. M. (1969): The structure of well-being, Chicago: Aldine.

Bradburn, N. M. (1983): Response e*ffects,* in: P.H. *Rossi* et al. (Hg.), Handbook of survey research, New York: Academic Press.

Carmines, E. G. & *Zeller,* R. A. (1979): Reliability and Validity Assessment, Beverly Hills, CA: Sage.

Converse, P. E. (1964): The nature of belief Systems in mass politics, in: *Apter,* D. E. (Hg.), Ideology and Discontent, Glencoe: Free Press.

Converse, P. E. (1970): Attitudes and non-attitudes: continuation of a dialogue, in: *Tufte,* E. R. (Hg.), The Quantitative Analysis of Social Problems, Reading/Mass.: Addison Wesley.

Duncan, O. D. (1984a): Measurement and structure: Strategies for the design and analysis of subjeetive survey data, in: *Turner,* C. F. & *Martin,* E. (Hg.), Surveying Subjeetive Phenomena, Vol. *2,* New York: Sage.

19 Eine der vermuteten Auswirkungen von 'hidden non-attitudes' ist, daß Sinnzusammenhänge von Items einer Skala erahnt und entsprechend konsistente und im Sinne von sozialer Erwünschtheit 'richtige' Antwortmuster produziert oder besser: fabriziert werden, obwohl eine auskristallisierte Meinung oder Einstellung nicht besteht (vgl. z.B. *Phillips* & *Clancy* (1972); *Schuman* & *Presser* (1978)).

Duncan, O. D. (1984b): Rasch measurement: Further examples and discussion, in: ***Turner,*** **C. F. &** ***Martin,*** E. (Hg.), Surveying Subjective Phenomena, Vol. 2, New York: Sage.

Jöreskog, **K. G. &** ***Sörbom,*** D. (1981): LISREL. Analysis of linear slructural relationships by maximum likelihood and least Squares methods, Chicago: National Educational Resources.

Lettau, F. (1987): RRM (Rating-Response-Modell). Ein PASCAL-Programm zur Maximum-Likelihood-Schätzung eines generalisierten logistischen Responsemodells für die Anwendung auf Items mit Rating-Antwortformaten, Beiträge IS/TUB 7, Institut für Soziologie, TU Berlin.

Lettau, F. (1989): Latente Variablen und Rating-Skalen, Berlin: Technische Universität Berlin.

Lumsden, J. (1977): Person reliability, Applied Psychological Measurement, 1, 477-482.

Lumsden, **J.** (1978): Tests are perfectly reliable, British Journal of Mathematical and Statistical Psychology, 31, 19-26.

Masters, **G. N. &** ***Wright,*** B. D. (1984): The essential process in a family of measurement models, Psychometrika, 49, 529-544.

Mortimer, **J. T. &** ***Lorence,*** J. (1979): Occupational experience and the selfconcept: A longitudinal study, Social Psychological Quarterly, 42, 307-323.

Perline, **R.,** ***Wright,*** **B. D. &** ***Wainer,*** H. (1979): The Rasch model as additive conjoint measurement, Applied Psychological Measurement, 3, 237-256.

Pfaff, **H.** (1989): Streßbewältigung und soziale Unterstützung, Weinheim: Deutscher Studicnverlag.

Phillips, **D. L. &** ***Clancy,*** **K. J.** (1972): Some effects of social desirability in survey studies, American Journal of Sociology, 77, 921-940.

Quinn, **R. P. &** ***Staines,*** G. L. (1979): The 1977 Quality of Employment Survey, Survey Research Center, University of Michigan/Ann Arbor.

Rasch, G. (1960): Probabilistic models for some intelligence and attainment tests, Kopenhagen: Danmarks Paedagogiske Institut.

Rasch, G. (1966): An individualistic approach to item analysis, in: ***Lazarsfeld,*** P. F. & ***Henry,*** N. W. (Hg.), Readings in Mathematical Social Science, Chicago: Science Research Associates.

Reuband, K.-H. (1990): Meinungslosigkeit im Interview, Zeitschrift für Soziologie, 19, 428-443.

Rosenberg, M. (1965): Society and the adolescent self-image, Princeton/NJ: Princeton University Press.

Rost,]., Davier, **M.,** ***Werthen,*** **M. &** ***Schütt,*** **A.** (1990): POLYRA-Polytomes Raschmodell, Institut für Pädagogik der Naturwissenschaften (IPN), Universität Kiel.

Schuman, **H. &** ***Presser,*** S. (1978): The assessment of 'no opinion' in attitude survcys, **in:** ***Schuessler,*** **K.** F. (Hg.), Sociological Methodology, San Francisco: Jossey Bass.

Smith, T. **W.** (1984): Nonattitudes: A review and evaluation, **in:** ***Turner,*** **C. F. &** ***Martin,*** E. (Hg.), Surveying Subjective Phenomena, Vol. 2, New York: Sage.

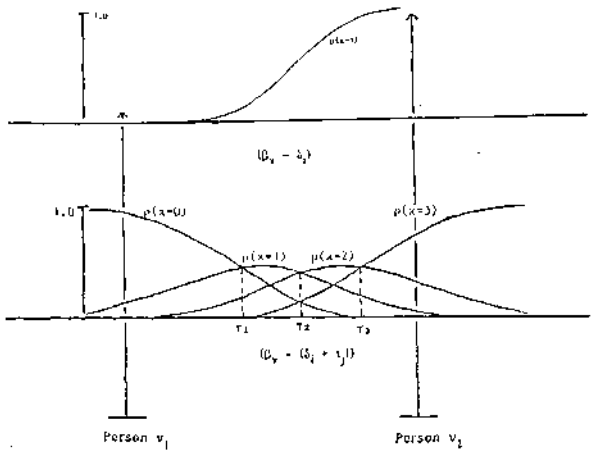
Sudman, **S. &** ***Bradburn,*** N. (1974): Response effects in survcys, Chicago: Aldine.

Waltz, **M.** (1987): Bedeutung der Familie bei der Infarktbewältigung, in: ***Badura*** et al. (Hg.), Leben **mit** dem **Herzinfarkt, Berlin:** Springer.

Wright, B. D. (1984): Additivity in psychological measurement, Research Memorandum #33, Chicago: **MESA** Psychometric Laboratory, University of Chicago.

Wright, B. D. & *Douglas*, G. A. (1977): Best procedures for sample-free item analysis, Applied Psychological Measurement, 1, 281-294.
Wright, B. D. & *Masters*, G. N. (1982): Rating scale analysis, Chicago: University of Chicago/MESA Press.
Wright, B. D. & *Panchapakesan*, N. (1969): A procedura for sample-free item analysis, Educational and Psychological Measurement, 29, 23-48.
Wright, B. D. & *Stone*, M. H. (1979): Best test design, Chicago: University of Chicago/MESA Press.

Anhang



$$p(x_{vi} = 1) = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}}$$

$$p(x_{vi} = 0,1,2,\cdots,k) = \frac{e^{\sum_{j=0}^{x_{vi}} (\beta_v - (\delta_i + \tau_j))}}{\sum_{k=0}^{m} e^{\sum_{j=0}^{k} (\beta_v - (\delta_i + \tau_j))}}$$

Nebenbedingungen $\tau_0 = 0$ und $\exp \sum_{j=0}^{n} (\beta_v - (\delta_i + \tau_j)) = 1$.

Möglicher Verlauf der modellimplizierten Antwortwahrscheinlichkeiten für ein Item i mit einem Itemparameter δ_i: Oben als dichotomes Item (= SLM); unten als Rating-Item (= RRM). Die τ-Parameter des RRM markieren die Wahrscheinlichkeits-Übergänge von einer Ausprägung zur nächsten.